NavGPT: Explicit Reasoning in Vision-and-Language Navigation with Large Language Models

Year: 2023 | Citations: 252 | Authors: Gengze Zhou, Yicong Hong, Qi Wu

---

Abstract

Trained with an unprecedented scale of data, large language models (LLMs) like ChatGPT and GPT-4 exhibit the emergence of significant reasoning abilities from model scaling. Such a trend underscored the potential of training LLMs with unlimited language data, advancing the development of a universal embodied agent. In this work, we introduce the NavGPT, a purely LLM-based instruction-following navigation agent, to reveal the reasoning capability of GPT models in complex embodied scenes by performing zero-shot sequential action prediction for vision-and-language navigation (VLN). At each step, NavGPT takes the textual descriptions of visual observations, navigation history, and future explorable directions as inputs to reason the agent's current status, and makes the decision to approach the target. Through comprehensive experiments, we demonstrate NavGPT can explicitly perform high-level planning for navigation, including decomposing instruction into sub-goals, integrating commonsense knowledge relevant to navigation task resolution, identifying landmarks from observed scenes, tracking navigation progress, and adapting to exceptions with plan adjustment. Furthermore, we show that LLMs is capable of generating high-quality navigational instructions from observations and actions along a path, as well as drawing accurate top-down metric trajectory given the agent's navigation history. Despite the performance of using NavGPT to zero-shot R2R tasks still falling short of trained models, we suggest adapting multi-modality inputs for LLMs to use as visual navigation agents and applying the explicit reasoning of LLMs to benefit learning-based models. Code is available at: https://github.com/GengzeZhou/NavGPT.